MULTIPLE PURPOSE OPTIMUM ALLOCATION IN STRATIFIED SAMPLING

H. O. Hartley, Texas A&M University

1. Optimum allocation in stratified sampling*

Consider a finite population of N units subdivided into H strata containing N_h units ($h = 1,2,\ldots,H$) respectively. Denote by y_{hi} the y of the i^{th} unit in the stratum h and by

$$Y_h = \sum_{i=1}^{N_h} y_{hi} \text{ and } \bar{Y}_h = Y_h/N_h \qquad (1)$$

the strata totals and means and by

$$Y = \Sigma Y_h \text{ and } \bar{Y} = Y/N \qquad (2)$$

the population total and mean. A random sample of n_h units is drawn at random from the h^{th} stratum and denote by

$$y_h, \bar{y}_h \qquad (3)$$

the sample strata totals and means corresponding to (2). The customary unbiased estimators of $\bar{Y}_h$ and $\bar{Y}$ are respectively given by

$$\bar{y}_h \text{ and } \hat{\bar{y}} = \Sigma (N_h/N) \bar{y}_h \qquad (4)$$

and the variance of $\hat{\bar{y}}$ by

$$\text{Var } (\hat{\bar{y}}) = \sum_h (N_h/N)^2 S_h^2 (1/n_h - 1/N_h) = \sum_h (N_h/N)^2 S_h^2/n_h - V_{con} \qquad (5)$$

where the h^{th} stratum variance, S_h^2, is given by

$$S_h^2 = (N_h - 1)^{-1} \sum_{i=1}^{N_h} (y_{hi} - \bar{Y}_h)^2 \qquad (6)$$

and V_{con} does not depend on the n_h.

If the cost of drawing the sample is given by the linear cost function

$$C_{TOT} = C_0 + \Sigma C_h n_h = C_0 + C \qquad (7)$$

then the classical 'optimum allocation' is defined as that set of n_h which minimizes Var $(\hat{\bar{y}})$ for a given cost C. From classical Lagrangean calculus we obtain

$$n_h^* = C (S_h N_h C_h^{-\frac{1}{2}}) / \sum_h (S_h N_h C_h^{+\frac{1}{2}}) \qquad (8)$$

resulting in a minimum variance of

$$V_{min} = C^{-1} N^{-2} (\sum_h S_h N_h C_h^{+\frac{1}{2}})^2 - V_{con} \qquad (9)$$

The formal solution (8) will of course, only be of practical use if

$$1 \leq n_h^* \leq N_h \qquad (10)$$

and will, in general, be fractional.

It will be shown in 2. that (8) does indeed yield an absolute minimum of (5) at constant cost.

*See e.g. Cochran, W. C. (1962)

2. Multiple purpose optimum allocation

Most sample surveys are concerned with obtaining estimates of a fairly large number of population parameters and not just the single linear estimate $\hat{\bar{y}}$ of $\bar{Y}$. Usually a large number of variables is measured for each sampled unit and not only is it required to estimate the population means for each of these but if the data are used in 'analytic studies' it may be of interest to estimate differences between all or some of the strata means for some or all of the variables but also for other subsections of the population called 'domains of study.' We propose to consider therefore a number of J different estimators $\hat{\bar{y}}_j$ which are linear functions of some or all of the strata means $\bar{y}_h$ and may involve these means for one or several of the variables. The variance of such linear estimators will be of the form

$$\text{Var } (\hat{\bar{y}}_j) = \sum_h a_{hj} (1/n_h - 1/N_h) = \sum_h a_{hj} n_h^{-1} - V_j \qquad (11)$$

where V_j does not depend on n_h. We retain the linear cost function (7) and consider three possible definitions of optimizations

(A)* Minimize a weighted sum of the J variances

$$\sum_{j=1}^{J} W_j \text{ Var } (\hat{\bar{y}}_j) = \sum_h n_h^{-1} \sum_{j=1}^{J} W_j a_{hj} - \sum_j W_j V_j = \sum n_h^{-1} A_{hj} - V \qquad (12)$$

at constant cost C.

*See Yates, F. (1953) and Cochran, W. G. (1962)

(B) Prescribe values v_j for the variances in the form

$$\sum_{h=1}^{H} n_h^{-1} a_{hj} = v_j \; : \; j = 1,2,\ldots,J \qquad (13)$$

and minimize the cost C given by (7) subject to (13).

(C)**Set tolerances for all variances in the form

$$\sum_{h=1} n_h^{-1} a_{hj} \leq v_j \qquad (14)$$

and minimize the cost (7) subject to the inequality restrictions (14). At first sight it may be argued that (B) is not necessary in view of (C) since one would surely not wish to force the variances to attain the upper tolerance v_j if it is possible to achieve smaller variances at the same cost. However, the utility of (B) lies in using its solution for variable v_j, under certain circumstances.

3. The solution to problem (A)

The solution to (A) is, of course, equivalent to a single purpose optimization with the minimum weighted variance (12) attained for

$$n_h^* = C(A_h/C_h)^{\frac{1}{2}} \Big/ \sum_h (A_h C_h)^{\frac{1}{2}} \qquad (15)$$

where

$$A_h = \sum_{j=1}^{J} W_j a_{hj}$$

and given by

$$V_{min} = C^{-1} \left(\sum_h (A_h C_h)^{\frac{1}{2}}\right)^2 - V \qquad (16)$$

In the special case of

$$W_{j'} = 1, \; W_j = 0 \text{ for } j \neq j' \qquad (17)$$

The problem reduces to the single purpose minimization of Var $(\hat{\bar{y}}_{j'})$ and (15) becomes

$$n_h(j') = C(a_{hj'}/C_h)^{\frac{1}{2}} \Big/ \sum_h (a_{hj'} C_h)^{\frac{1}{2}} \qquad (18)$$

leading to the minimum

$$V_{min}(j') = C^{-1} \left(\sum_h (a_{hj'} C_h)^{\frac{1}{2}}\right)^2 - V_{j'} \qquad (19)$$

**See Dalenius, T. (1957), Yates, F. (1953) and Cochran, W. G. (1962)

We now show that (19) is an absolute minimum for Var $(\hat{\bar{y}}_j)$:

Consider an allocation n_h satisfying the given cost condition

$$C = \sum_h C_h n_h \qquad (20)$$

then we have to show that the variance of $\hat{\bar{y}}_j$ computed from (11) and using the sample sizes n_h will exceed $V_{min}(j)$. Using (11), (19) and (20) we obtain

$$C(\mathrm{Var}(\hat{\bar{y}}_j) - V_{min}(j)) =$$

$$= \left(\sum_h C_h n_h\right)\left(\sum_h a_{hj} n_h^{-1}\right) - \left(\sum_h (a_{hj} C_h)^{\frac{1}{2}}\right)^2$$

$$= \sum_h C_h n_h \left((a_{hj}/C_h)^{\frac{1}{2}} n_h^{-1} - Av\right)^2 \geq 0 \qquad (21)$$

where Av is the weighted average.

$$Av = \sum C_h n_h (a_{hj}/C_h)^{\frac{1}{2}} n_h^{-1} \Big/ \sum C_h n_h \qquad (22)$$

Formula (21) shows that the minimum variance is attained if and only if the n_h satisfy (13) and will in general provide the amount by which Var $(\hat{\bar{y}}_j)$ exceeds $V_{min}(j)$.

4. The solution to problem B

The J linear equations (13) for the H variable n_h can, of course, only be satisfied if H-J of the equations are linearly dependent upon H of them and even then the solutions may not yield positive n_h. We shall therefore confine our discussion of this problem to such specifications of v_i which are 'of interest' that is to situations in which the system (13) has at least a one parametric infinite set of solutions. A necessary condition for a minimum of the cost C under the restrictions (13) is given by (15) where the weights W_j are now to be interpreted as Lagrangean multipliers and must be determined by substituting (15) into (13). It is easy to show by reference to the 2nd order differentials that (15) is also a sufficient condition for the n_h^* to yield an absolute minimum of the cost C provided the W_j are determined to satisfy (13). In practice, however, one would not proceed in this manner but rather start from the Lagrangean weights W_j, and go through the following steps:

(i) Choose weights W_j representing the relative importance of the variances $V(\hat{y}_j)$ and fix a budget C for the survey. Solve problems (A) yielding the optimum allocation n_h^* given by (15).

(ii) By substituting the n_h^* in (15) compute individual variances $v_j - V_j$ for the estimators $\hat{y}_j$. Since the n_h^* are now also the solutions to problem (B) it can be stated that at least these variances $v_j - V_j$ can not be achieved at a smaller cost than the budget C.

(iii) Compare the $v_j - V_j$ with the $V_{min}(j)$ for the same budget given by (19) and increase the weights W_j for such $\hat{y}_j$ where the excess is 'disappointingly large.'

(iv) If the adjustment in (iii) does not lead to a satisfactory set of v_j and if a constant percentage decrease in the v_j is desired the corresponding percentage increase in the budget C will achieve this.

There are obvious limitations to the formulation and solution of the multiple purpose optimization problem in the form (A) and (B), and we may summarize them as follows:

The main reservation about minimizing a weighted variance (as in (A)) puts the onus on the choice of weights W_j which may result in unreasonably high variances of some of the $\mathrm{Var}(\hat{y}_j)$ in the weighted sum. The approach in (B) however, does much to rectify this disadvantage: It not only shows that at least the actually attained $\mathrm{Var}(\hat{y}_j)$ have been met with minimum cost, but it also gives a feed-back for the improvement of the choice of weights. There remain, however, two main disadvantages. First, the procedure described above will in general require that $J<H$, i.e. that the number of estimators entering into the optimization does not exceed the number of strata, and moreover should be moderate or small for convenience in the adjustment of the W_j. Secondly, the solution n_h^* may well exceed N_h and we have so far not discussed what to do in such situations. All these problems can be resolved if we adopt the formulation (C) and solve it by non-linear programming.

5. The solution of problem C by non-linear programming

In finding the minimum cost C under the inequality restrictions (14) we find it convenient to introduce the reciprocals

$$r_h = 1/n_h - 1/N_h \quad h = 1,\ldots,H \tag{23}$$

as the elements of our activity vector r which results in a convex activity space with the linear boundaries defined by

$$Ar \leq v - V \tag{24}$$

and the upper 'bounds'

$$0 \leq r_h \leq 1 - 1/N_h \tag{25}$$

where A is the H x J matrix of the a_{hj} and $v - V$ the J-vector with elements $v_j - V_j$ that is the set of tolerances for the variances $\mathrm{Var}(\hat{y}_j)$.

No assumptions need be made concerning the rank of A or the magnitude of H and J except within the framework of available computer codes. The 'objective function,' i.e. the cost now becomes the convex function

$$C = \sum_{h=1}^{H} C_h \left(r_h + \frac{1}{N_h}\right)^{-1} \tag{26}$$

and is of a form described as 'separable' see Charnes and Lemke (1954) and Hartley (1960). This fact would make available the procedure by Hartley (1960) which would involve an approximation to the hyperbolae $\left(r_h + \frac{1}{N_h}\right)^{-1}$ by a moderate number of linear line segments which method has been shown to reduce the problem to linear programming. Alternatively the method recently published by Hartley and Hocking (1963) could be used which does not require polygonal approximations. For the details of the algorithm we must refer to this paper. We confine ourselves here to stating that a new variable is introduced in the form

$$r_{H+1} = -C = -\Sigma\, C_h \left(r_h + \frac{1}{N_h}\right)^{-1} \tag{27}$$

and that r_{H+1} is maximized whilst (27) occurs as a (non-linear) restriction. The problem is solved in the dual form which leads to the following tableau.

Tableau I. Dual tableau for non-linear programming problem

	j=0	j= 1,...J	h= 1,...h=H	h= H+1	h= H+2
h=0	0	$v_j - V_j$	$(1-1/N_h)$	M	C^+
h=1	0		-1	0	
	.		-1	.	
	.		.	.	
	.		.	.	
	.		.	.	
	.	$-a_{hj}$	.	.	$+C_h(r_h+\frac{1}{2}N_h)^{-2}$
	.		-1	.	
h=H	0		-1	0	
h=H+1	1	0...0	0 . . . 0	1	1

Tableau I will be recognized as the dual tableau in standard form, for maximizing r_{H+1} subject to the restrictions $Ar \leq v-V$ and $r_h \leq (1-1/N_h)$ except that the line h=0 represents the negative of the dual objective function and that the last two columns require some explanation: The column h=H+1 is an 'artificial vector' to supplement the (H+1) x (H+1) identity matrix of slack vectors (not shown in Tableau I) for an initial "basis." Its 'penalty' M will eventually drive it out of the basis. The last column represents the non-linear restriction (27) and is non-standard. Whilst for an explanation of this column we must refer to Hartley and Hocking (1963) we should state here that its first element is given by

$$C^+ = -\Sigma\, c_h(r_h+\tfrac{1}{2}N_h)^{-1} - \Sigma\, c_h(r_h+\tfrac{1}{2}N_h)^{-2} r_h \quad (28)$$

and that it is evaluated for varying argument vectors r_h in the course of the simplex process and may contribute several columns for the current basis matrix.

It will be noted that the problem leads to a dual Tableau of size (H+2) x (J+H+3) which is quite a feasible size for a high-speed computer even if H is of the order 50 and J of the order 200. Three small numerical examples are given in Hartley and Hocking (1963).

It must of course be remembered that the algorithm of the non-linear programming technique only yields a numerical optimum allocation n_h^* (in the r_h^* scales) for the particular problem and no general formula for the n_h^*. It may therefore rightly be asked whether there are techniques available which exhibit numerically the effect on the r_h^* of altering the specified variance tolerances $v_j - V_j$. There is indeed such a technique available which is known under the name of 'parametric programming' and which is incorporated in most computer codes. Another question which may be asked concerns the uncertainty in the a_{hj} which depend on the strata variances. Since the strata variances S_h^2 would normally not be known but estimated one may wish to regard the a_{hj} as stochastic variables. Such a model would lead to methods of stochastic programming. Here we are certainly more restricted with regard to the availability of methods and computer codes.

More recently we have obtained some new results on convex parametric programming using a modification of Hartley and Hocking (1963) which will be published shortly. With these methods it will be possible to examine how an alteration of the variance tolerances $v_j - V_j$ effects the optimum allocation n_h^* and the cost C.

References

Charnes, A. and Lemke, C. E. (1954), 'Minimization of Non-linear Separable Convex Functionals.' Naval Research Logistics Quarterly, I (1954), 301-12.

Cochran, W. G. (1962), 'Sampling Techniques' 2nd Edition J. Wiley and Sons, New York.

Dalenius, T. (1957), 'Sampling in Sweden. Almquist and Wicksell, Stockholm.

Folks, J. Leroy and Antle, C. E. (1965), 'Optimum Allocation of Sampling Units to Strata When There are R Responses of Interest,' JASA, 1965, p. 225-233.

Hartley, H. O. (1961), 'Non-Linear Programming by The Simplex Method.' Econometrica, 29 223-237.

Hartley, H. O. and Hocking, R. (1963), 'Convex Programming by Tangential Approximation.' Management Science (in Press)

Kokam, A. R. (1963), 'Optimum Allocation in Multivariate Surveys.' JRSS, Series A, 126 p. 557-565.

Yates, F. (1955), 'Sampling Methods for Censuses and Surveys.' 2nd Edition Hafner, New York.